**Manuscript ID** remotesensing-3458794

**Title** MosaicFormer: A Novel approach for Remote Sensing Spatio-temporal Data Fusion for lake water monitor

*Remote Sensing*

Dear Editor:

We sincerely thank you for your insightful and constructive comments. We also greatly appreciate the editor’s invitation to submit a revised version of the manuscript. In response to the reviewers’ suggestions, we have thoroughly revised the manuscript. Below, you will find our point-by-point responses in blue text following the reviewer comments. All comments have been carefully considered and addressed in the revised version. We believe that your concerns have been fully resolved, but please do not hesitate to reach out if there are any further questions or suggestions.

Thank you for your consideration. I look forward to hearing from you.

Yours Sincerely,

Dr. Aifeng Lyu

Professor, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China

Physical address: 11A, Datun Road, Chaoyang District, Beijing, 100101, China

Phone: 010-64856539

Email: lvaf@igsnrr.ac.cn

Comments in blue and our response in black

**Reviewer #1:**

***Comments:***R1C1: The overall content of the article is unclear, and it is not clear what the focus of the article is. Is it focused on the research of spatiotemporal fusion methods or the evaluation and analysis of lake water monitoring? If it is the former, then the article has not made comparisons in terms of methodology, no ablation experiments, and no demonstration of the superiority of the proposed method. If it is the latter, the study only presents an analysis figure (Figure 5) but does not highlight the differences between lake area monitoring using MODIS and the fused results. Whether discussing methodology or application analysis, the descriptions are superficial and lack detail, failing to showcase any innovation

**Response:** Thank you for your suggestion.

这个要说侧重点在前面，后面只是一个实际应用的案例。

按照审稿人的意见：

* 与多个方法进行比较，尤其是Swin transformer相关的
* 添加消融实验，包括Swin Transformer模块、MAE模块、VQ-VAE模块
* 添加更多的技术细节，创新性的描述

创新性如下：

* 新模型的引入本身就是一种创新
* 概念上的创新：less is more

R1C2: In the introduction part, the authors summarize previous work, but does not clearly point out their shortcomings or how the proposed method addresses these issues.

**Response:** Thank you for your insightful comments.

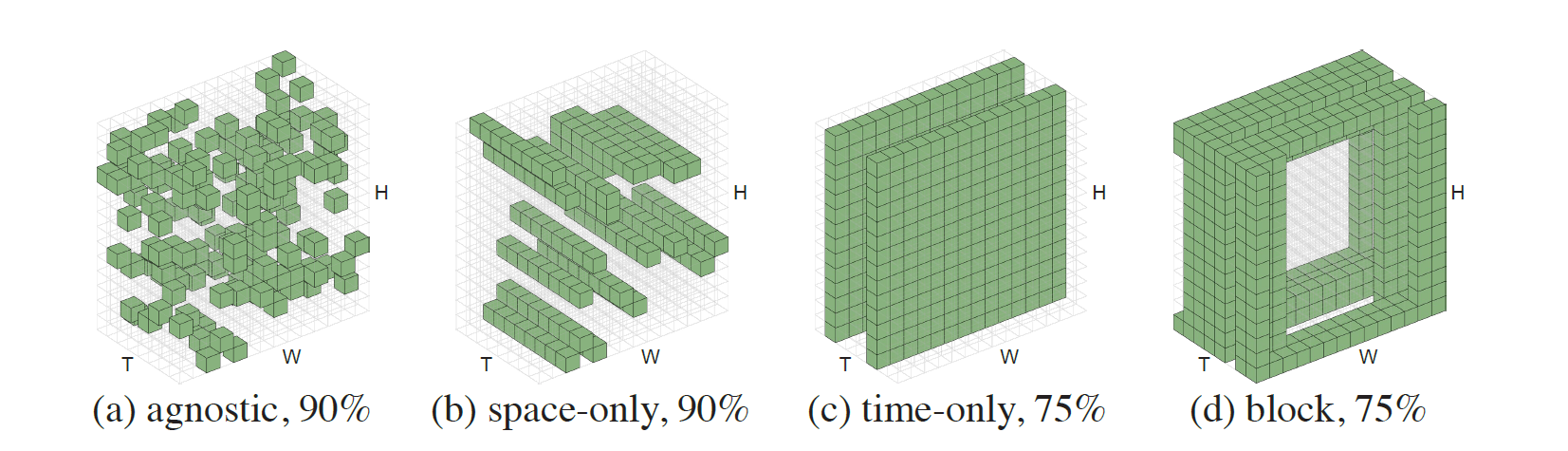
丰富引言

Additionally, the subsequent methodology section does not highlight the advantages or innovative aspects of the proposed approach. In the final paragraph of the introduction, the authors mention combining Masked Autoencoders and Transformers and introduce the concept of "less is more," emphasizing that Masked Autoencoders randomly mask out some low-frequency information. However, I believe that this low-frequency information (mainly from MODIS data) contains important temporal variations. If low-frequency information is randomly masked, wouldn’t this remove some of the crucial temporal changes? The validity of this concept and its applicability to spatiotemporal fusion need to be supported by relevant references or experimental evidence.

**Response:** Thank you for your insightful comments.

这个是创新性之一，为了回复他的担忧，我们使用了图b类似的思路，但并非完全一样，具体来说是选择一些连续的时间点，使之在训练阶段永远不会被mask。正确的方案是a+b，说在模型方法描述中添加了更多细节，当然也包括重新绘制下面这种概念图。

Feichtenhofer, C., Li, Y., & He, K. (2022). Masked autoencoders as spatiotemporal learners. Advances in neural information processing systems, 35, 35946-35958.



*R1C3: The methodology section is too brief and lacks detail. The paper mentions using MAE and Swin- Transformer, but Swin-Transformer has already been applied to spatiotemporal fusion in Chen et al. (2022). Their method includes a Feature Extraction Module (FEM) and a Multi-Level Fusion Module (MFM). If the only innovation here is adding MAE as an encoder, it is insufficient. Additionally, the authors mention that the model contains an improved MAE encoder and a vector quantized variational autoencoder (VQ-VAE), but how was the MAE improved? Is the improved MAE encoder superior to the original version? Was pre-trained weight used? These details are unclear. If only a module is added on the basis of Swin-Transformer, such innovation is not enough for publication in this journal.*

*Reference: Chen G, Jiao P, Hu Q, et al. SwinSTFM: Remote sensing spatiotemporal fusion using Swin Transformer [J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-18.*

**Response:** Thank you for your insightful comments, which are crucial for clarifying our model's description.

首先需要在方法部分添加更多细节，这里我来加。

其次要突出创新性，包括我们的概念less is more，对传统的Swin-Transformer改进

然后添加消融实验

其实超分辨率或融合还是扩散模型比较能打（需要你找找相关文献支持），Swin-Transformer参数量大，融合训练数据也不多，其实这种模型表现不好。我们做出了比较好的改进。

*R1C4: The results analysis section only presents a comparison between the fused results and the ground truth, but the descriptions are vague and lack details. There is no comparison with other methods, nor any ablation experiments to demonstrate the superiority of the approach.*

**Response: 首先这里添加细节，与其它方法的比较。**

**额外添加消融实验（之前提过了）**

*Recently, Professor Gong Peng from Tsinghua University released global seamless daily data from 2000 to 2022. Whether the results of this paper are superior to the published data in terms of details or regional aspects? If the published data are used, can daily monitoring of lake areas be carried out? Then where lies the superiority of the method in this paper?*

**Response: 这里需要你下载宫鹏的数据进行对比。希望效果能优于他**

**他的数据当然也可以达到每日监测，但如果针对特定区域，我们的模型效果要更好，不排除这是宫鹏的学生审的，因此要夸他们模型的同时引用相关文献**

**我们的模型可以融合Landsat和sentinel（只需要改一下输入和输出的维度），能达到10m（这里需要copy一些之前GEE CNN那个结果展示到这里）**

*Additionally, in Figure 4b, the title refers to a "super-resolved 30m image." While spatiotemporal fusion contains the idea of super resolution, but it is different from remote sensing image super resolution. I think it is more appropriate to use spatiotemporal fusion.*

**这里描述相应修改**

*Furthermore, in Figure 4，b, c, e, and f, the fused results have significant color differences from the ground true images. Why is there such a large radiometric discrepancy? In theory, maintaining radiometric or spectral consistency is crucial in spatiotemporal fusion. If the radiometric values differ significantly, it will introduce large errors in applications.*

**我来解决这个色差问题，可能是因为训练epoch不够**

*Additionally, in Figure 3, while different methods are compared, it is unclear whether these methods are representative and state-of-the-art. The scatter plot (Figure 3b) also lacks clear explanations of the x- and y-axes.*

*Professor Gong Peng's article and data can refer to the following papers:*

*Chen S, Wang J, Gong P. ROBOT: A spatiotemporal fusion model toward seamless data cube for global remote sensing applications[J]. Remote Sensing of Environment, 2023, 294: 113616.*

*Chen, S., Wang, J., Liu, Q., Liang, X., Liu, R., Qin, P., Yuan, J., Wei, J., Yuan, S., Huang, H., and Gong, P.: Global 30 m seamless data cube (2000–2022) of land surface reflectance generated from Landsat 5, 7, 8, and 9 and MODIS Terra constellations, Earth Syst. Sci. Data, 16, 5449–5475 2024.*

**Response:** Thank you for your comment, which raises a crucial point.

*R1C5: In lake monitoring (Figure 5b), the fused image has not only large tonal differences from the ground true images but also exhibits Landsat 7 strips. These stripes should not appear in the fusion results if there is sufficient training data. The authors should investigate the cause of these artifacts. In the dataset released by Professor Gong Peng, there are no such stripes. More importantly, if we have Professor Gong Peng's data, then where lies the innovativeness and pioneering in this paper?At the same time, this study emphasizes the importance of lake monitoring in the introduction. Then, the spatio-temporal fusion method in the article should serve daily lake monitoring. Where does the method reflect this point? In my opinion, the method is more general and not specifically designed for the characteristics of lake monitoring.*

**Response:** Thank you for pointing out the inaccuracies in lines 90-91 of our manuscript.   
条带问题是因为训练数据不足导致的（仅有17年的10对影像？而且有云），该区域属于稀缺资料地区（你下载一下宫鹏的数据看看这里）

创新性可以说模型比较新，在区域训练能得到更好的结果

该方法确实是一个通用模型，适合在各种区域进行监测，这只是一个例子

*R1C6: The paper also has formatting and detail issues. For instance, abbreviations should be defined when first mentioned. In line 125, "Masked Autoencoders" appears without a full English definition. Additionally, many references are outdated, mostly from around 2020 or earlier. While reviewing traditional methods is necessary, the paper lacks discussions of recent advances in deep learning- based spatiotemporal fusion. The authors should compare their method with more recent state-of-the-art models rather than only with traditional approaches.*

**Response:** Thank you for your insightful comments.

调查一下最新的融合和超分辨率模型

**Reviewer #2:**

***Comments:***

*R2C1: Consider incorporating other remote sensing data sources (such as Sentinel-2) to further improve the accuracy and robustness of data fusion.*

**Response:** Thank you for your constructive feedback on the introduction section.

这个把之前的CNN结果拿过来一部分吧

*R2C2: The current study only uses Hala Lake as a case study. It is recommended to validate the method on a wider range of water bodies (such as urban lakes, plateau lakes, river reservoirs, etc.) to demonstrate its general applicability.*

**Response:** Thank you for your insightful comment, which highlights a crucial issue.

这个把之前的CNN结果拿过来一部分吧

*R2C3: For MODIS imagery, if a pixel corresponds purely to the lake, the accuracy should be high. However, if the pixel lies on the lake boundary, being a mixed pixel, the proportion of the lake within the pixel may vary, potentially leading to the learning of incorrect features. How does the author address this issue?*

**Response:** 这个只能加讨论，如果湖泊边界变化不明显（干枯不分明），效果较好；反之

**Reviewer #3:**

***Comments:***

*R3C1:*

**Response:** Thank you for your comments and for providing references to related works.

*R3C2:*

**Response:** We appreciate your feedback concerning the performance issues in arid regions and human-influenced areas, as indicated in Figures 5 and 6.

*R3C3:*

**Response:** Thank you for your suggestions.

*R3C4:*

**Response:** Thank you for your feedback regarding the description of existing studies and the first category in our manuscript.

*R3C5:*

**Response:** Thanks! We have modified that.

*R3C6:*

**Response:**